Facial Expression Recognition Based on Multi-dataset Neural Network

Bin YANG, Zhenyu LI, Enguo CAO

School of Design, Jiangnan University, Lushan road 1800, Wuxi, China

Yangbin@jiangnan.edu.cn, {150245114, 24942849, 627224124}@qq.com

Submitted April 9, 2019 / Accepted December 16, 2019

Abstract. *Facial activity is the most powerful and natural means for understanding emotional expression for humans. Recent years, extensive efforts have been devoted to facial expression recognition by using neural networks. However, automated emotion recognition in the wild from facial images remains a challenging problem. In this paper, an effective facial expression recognition scheme is proposed. A multi-dataset neural network is developed to learn facial expression features in several different but related datasets. The novel multi-dataset network fuses the intermediate layers of a deep convolutional neural network (CNN) by using separate CNNs and a multi-dataset loss function. Experimental results performed on emotion database demonstrate that our proposed method outperforms state-of-the-art.*

Keywords

Facial expression recognition, design of network architecture, deep learning, human–computer interaction, convolutional neural network

1. Introduction

Facial Expression Recognition (FER) is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life. A core of social and emotional intelligence is to notice and understand emotional states and other social signals of a person. An automatic facial expression recognition system is desired in emerging applications in human-computer interaction (HCI), such as online/remote education, interactive games, and intelligent transportation.

Recent years, many works have been published to accurately recognize user behavior [1–3]. Support vector machine (SVM) is a supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It has been widely implemented for HCI. Although significant performance achieved by these SVM-based methods [1], major challenges remain open for HCI. A main concern is that approaches based on SVM may obtain unsatisfied results while performing on task of emotion recognition, particularly in facial expression recognition. Numerous researchers have been conducted on automatic FER due to its practical importance in human-computer interaction systems, such as user-interface design, medical treatment and driver fatigue surveillance.

These years, deep learning techniques have gained a good reputation in computer vision fields. Usually, the features learned via deep learning have better representations of the data than the handcrafted features [7]. Comparing to feature extraction based on expert's knowledge, deep learning is considered a more effected way to discover the discriminant representations inherent in human faces by incorporating the feature extraction into the task learning process. Deep learning can be used by nonexperts for their researches and/or applications, especially in FER field.

FER systems can be roughly divided into two main categories: statistic image FER and dynamic sequence FER [3]. In static-based methods, the facial features [4] are generated with only spatial information from the current single image. On the other hand, dynamic-based methods consider the temporal relation among contiguous frames in the input facial expression sequence [5]. In this work, we focus on the former one (static-based method) due to its relatively small computation. Multi-dataset learning is developed by human learning activities where people often take advantage of foreknowledge of previous tasks to help learn a new task. Compared with the current state-of-the-art of FER techniques, the contribution of our work can be summarized as follows:

- Many existing techniques are focusing on improving the performance in several public FER datasets, which may lead to a limitation in real-world case. The proposed neural network is trained by the images not only in FER datasets, but also in wild experiment datasets. Therefore, it can recognize expressions in real-life scenarios.
- A multi-dataset loss function is proposed to enhance the discriminative power of the deeply learned features.
- A novel multi-dataset network architecture is developed to improve the robustness of our proposed FER scheme.

DOI: 10.13164/re.2020.0259

The rest of the paper is organized as follows. Section 2 presents the review of FER approaches. The proposed method is presented in Sec. 3. Section 4 presents the experimental results and discussions. Finally, the concluding remarks are given in Sec. 5.

2. Related Works

Although, various approaches represent important results for FER, especially considering that the problems they tackle were previously (almost) unexplored. A large set of tools is now available. Among them, methods based on artificial neural networks were proven to be promising for exposing some challenge exemplars. CNN has been extensively used in diverse computer vision applications, including [6], steganography [7], FER and other applications [8–10]. Fasel [11] found that CNN is robust to face location changes and scale variations. He also discovered that CNN is more efficient than multilayer perceptron (MLP) in human face detection if previously face pose variations is missed. Matsugu et al. [12] proposed a CNN-based scheme to deal with the problems as well as translation, rotation, and scale invariance when recognizing the facial expressions.

In the wild cases, fusion techniques can be used to combine multimodal features, the accuracy of FER is unpleased due to the powerless of single descriptor. Sikka et al. [13] combined multiple visual descriptors and paralinguistic audio features to classify video clips in multimodal way. The extracted features were combined by using Multiple Kernel Learning. The visual and acoustic features were merged in their method. Girshick et al. [14] explored the usage of Regions with CNN features (R-CNN). High-capacity CNN was applied to the bottom-up region to localize and segment objects. Inspired by R-CNN, Sun et al. [15] designed two temporal-spatial dense scale-invariant feature transform (SIFT) features and combined these multimodal features to recognize human expression from image sequences. Linear SVM and partial least squares were used as classifiers for those kinds of features on the static and acted facial expression in the wild. They also proposed a fusion network to combine all the extracted features at the decision level. They gained significant scores on public datasets.

Many FER schemes seldom consider the interaction between potential factors. In wild scenarios, FER may be interfered by head pose, illumination and so on. On the other hand, multi-dataset learning models aim to learn a cross-dataset parameter sharing strategy, which reflects the similarities and differences between datasets. The goal of such selective parameter sharing is to make the difference robust when exploiting data from different datasets. Figure 1 is an illustration of multi-dataset network.

Since, the hyper-features must be associated in a way that effectively encodes common features for multiple datasets. Multi-dataset neural networks are generated in the presence of task data in several different but related datasets. The differences between multi-datasets learning and multi-task learning are subtle. Some multi-dataset learning problems can be solved by methods developed for multi-task. In order to learn different tasks using images from different data-bases, Fourure et al. [16] proposed a multi-task CNN for semantic segmentation of outdoor images. A selective soft-max cross entropy function was developed. By using such function, images from another task can be trained without parameter losing. They illustrated their capacity by the following example. Let database D_{Grass} and $D_{\text{Vegetation}}$ is the labeled database of grass and vegetation, respectively. The multi-task CNN in [16] can accurately estimate the image from D_{Grass} when the input image belonged to $D_{\text{Vegetation}}$. But for the human face, the accuracy of their approach is not satisfactory [17]. Zhang et al. [18] adopted a cascaded structure with three stages of carefully designed deep convolutional networks that predict face and landmark location in a coarse-to-fine manner. They achieved superior accuracy on several challenging dataset.

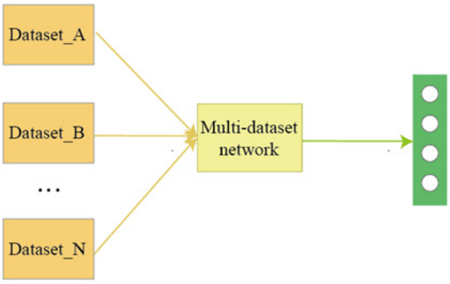

Fig. 1. An illustration of multi-dataset network.

To solve the problem of transferring representations learned from multiple source datasets, Xian et al. [19] utilized multiple convolutional neural network (CNN) models trained on different labelled source datasets by feeding soft labels obtained by clustering on target dataset to each other. The enhanced model can learn more discriminative person representations than the single model trained on multiple datasets. In [20], lightweight-but-powerful fully convolution network was proposed. A dense anchor strategy and a scale-aware anchor matching scheme were used to improve the recall rate of small faces. Recently, a four-task network was proposed to detect human face [21]. Facial landmark and pose can be located and estimated simultaneously. In addition, the method can estimate human gender. They developed a separate fusion-CNN to fuse the intermediate layer features. However, the overall dimension of these intermediate layer features is too large to be learned by network. The weakness of their approach is to use only a single data source. All images used for training must be labeled with all relevant tasks.

Inspired by the promoting works based on multi-dataset learning, we proposed a novel FER neural network. By using such network, knowledge could be transferred from other relevant tasks. In the meanwhile, the disturb factor would be disentangled.

3. Proposed Method

3.1 Design of Emotion Database

In order to meet the needs of different scenarios, we use two kinds of datasets to form an emotional database. Actually, we can use more than two datasets to improve the robustness of the proposed FER network. In this work, the experimental database contains two types of datasets: the facial emotion data in experiment and the facial expressions data in wild scenarios.

The former one includes FER2013 dataset [22] and experimental collected images. FER2013 is a large-scale and unconstrained database collected automatically by the Google image search API. All images have been registered and resized to 48 × 48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region. FER2013 contains 28709 training images, 3589 validation images and 3589 test images with seven expression labels. However, the number of Asian faces in many public datasets is quite small. To improve the robustness of our proposed method, especially for Asian faces, we added a large number of Asian faces to the experimental dataset. 1400 facial images with ages from twenty to fifty were obtained. Every people were requested to offer seven facial expressions (i.e. happiness, sadness, surprise, anger, disgust, fear, and neutral). A sample of our collection is shown in Fig. 2.

The second dataset includes the Static Facial Expressions in the Wild (SFEW) dataset [23] and experimental collected images. SFEW was usually used to support training data for emotion recognition in wild-scene. It consists of short video clips extracted from popular Hollywood movies. Each clip contains a film actor who has been labeled into one of the seven basic facial expression categories, namely Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. Several sample images in SFEW are presented in Fig. 3. Similar to the previous dataset, we added many faces of Asian actors to the second dataset.

As a pre-processing step, face alignment is necessary in many facial recognition applications. In this work, we use the face alignment strategy proposed in [24] to robustly align human face 'in-the-wild'. Finally, each aligned facial image was cropped into 48 × 48 pixels and was transformed to grayscale.

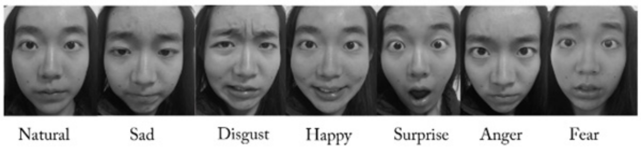

Fig. 2. A sample of our collected dataset for FER.

Fig. 3. Sample images in SFEW.

3.2 Deep Multi-dataset Network

Since the ability of computation is still limited especially in wearable equipment, multi-application would be more appropriate in some cases. The proposed work needs to adopt real-world accommodations in more convenient way. Many existing networks for FER focus on a single task and learn features that are sensitive to expressions without considering interactions among other latent factors. In the real world, different factors would distract the FER task, such as lighting, head pose, etc. Based on this, we used multi-dataset leaning technique to improve the robustness of FER networks.

3.2.1 Network Architecture

As shown in Fig. 4, the proposed architecture for facial expression recognition includes a pair of identical components CNN. The parameters *C* and *m* have different values for different layers in the network as is demonstrated in the figure.

Each component contains five convolutional layers. Following the final convolutional layer, two fully connected (FC) layers consisting of 2048 neurons are employed. In the convolutional layers, we use kernel size of $m \times m \times C$, where C is the depth of a filter and m is the size of convolutional kernel. After the convolutional layer there is the pooling layer. We used max-pooling technique to decrease the size of feature maps. Let i denote the index of a max-pooling layer. The layer's output is a set O_i of square maps with size w_l. We get the O_i from O_{i-1}. The square maps size w_l is obtained by $w_l = w_{l-1} / k$, where k is the size of the square max-pooling kernel.

As an extension of a Rectified Linear Unit (ReLU) [23] activation function, Concatenated Rectified Linear Units (CReLU) was proposed to reduce the number of computations by half without losing accuracy [24]. It conserves both positive and negative linear responses after convolution so that each filter can efficiently represent its unique direction. Because CReLU retains the available information of the input while keeping the non-saturated non-linearity, it may

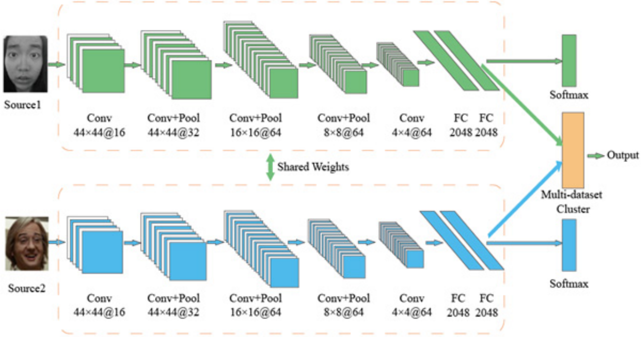

Fig. 4. The proposed multi-dataset CNN.

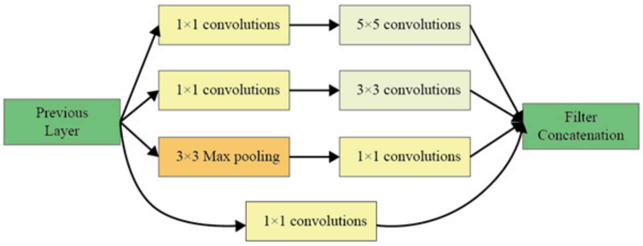

Fig. 5. The proposed inception module.

be beneficial to more complex machine learning tasks, such as structured output prediction and multitask learning.

To reduce the number of parameters and improve computational efficiency, an inception module [25] is used in our proposed network. Inception module produces output activations of different sizes of receptive fields. The layers closed to the input will bring the units to a local region. Therefore, we chose 1 × 1 kernel convolution layers after the second convolutional layer to capture substances that vary greatly in size. The architecture of the proposed inception module is presented in Fig. 5.

3.2.2 Network Integration

The output of each component network should be integrated in the final step. The common way is use weighted summation. Let S_i is the final score of emotion $i = 1,...,p$ where p is the total number of emotion class, S_i can be calculated as:

$$S_i = \sum_{j=1}^{N} \lambda_j C_{j,i}$$

where m is the number of component network (i.e. N is 2 in this work). The sum of the weights λ_j is 1, i.e. $\sum_{j=1}^{N} \lambda_j = 1$. In general, parameter λ_j is often set to the same value.

The above method is simple to use, but may not make full use of the ability of the two networks with different dataset as input. Therefore, a fine-tuning method with a novel loss function for alternating integration of two networks was proposed. By using this method, the performance of FER in multi-dataset can be improved.

Different loss functions for each task and train alternatively for the different domains were developed to handle such problem. Softmax loss technique is usually used in convolutional CNN to force the features of different classes staying apart. To improve the performance and further reduce the intra-class variations, Center loss method [26] is developed. It can estimate the center of each class of features, and drag the features belonged to the same center at the same time. We start with the definition of Center loss function L_C:

$$L_C = \frac{1}{2}\sum_{i=1}^{k} \left\| x_i - c_{y_i} \right\|^2 \tag{1}$$

where y_i and x_i are the class label of the i^{th} sample and the feature of the i^{th} sample generated from the fully-connected layer, respectively. c_{y_i} denotes the center of the cluster in which all samples are labeled as y_i, and k is the number of the samples. They compute the Joint supervision loss value to minimize the intra-class variations while keeping the features of different classes separable [26]. The centers will be updated in each iteration using Stochastic Gradient Descent (SGD) [27] as part of the CNN training.

During its forward propagation, the weighted sum of the softmax loss and the center loss is calculated:

$$L = L_S + \lambda L_C \tag{2}$$

where L_S is the softmax loss, and λ is a parameter to assign the weight of softmax loss and center loss.

In backward propagation process, the partial derivative of the center loss L_C with respect to the input sample x_i can be calculated as:

$$\frac{\partial L_C}{\partial x_i} = x_i - c_{y_i} \cdot \tag{3}$$

The centers are updated in the iterative optimization as defined below:

$$\Delta c_j = \frac{\sum_{i=1}^{k} \delta(y_i, j) \cdot (x_i - c_{y_i})}{1 + \sum_{i=1}^{k} \delta(y_i, j)} \tag{4}$$

where $\delta(y_i, j)$ is defined as:

$$\delta(y_i, j) = \begin{cases} 1, & y_i = j, \\ 0, & y_i \neq j. \end{cases} \tag{5}$$

However, in some cases, the clusters of different classes would be overlapped by using Center loss technique. To enhance the discriminative power of the deeply learned features, we proposed a novel multi-dataset (MD) loss function for FER. It can further improve the distance between different clusters.

Different to center loss, we calculate all the distances between a sample and other class centers. The objective is initially defined as follows:

$$L_{MD} = -\sum_{i=1}^{N} \log \frac{\exp\left(\frac{1}{\varphi} \left\| x_i - c_{y_i} \right\|_2^2 - \lambda \right)}{\sum_{j=1, j\neq i}^{k} \exp\left(-\frac{1}{\varphi} \left\| x_i - c_{y_i} \right\|_2^2 \right)} \tag{6}$$

where k is the number of clusters, and λ is a predefined margin parameter. λ can be calculated as:

$$\lambda = \sum_{i=1,\ j=1, j\neq i}^{k} \frac{c_i \cdot c_j}{\left\| c_i \right\|_2^2 \left\| c_j \right\|_2^2} \tag{7}$$

where c_i and c_j are the i^{th} and j^{th} center of the cluster, respectively. φ is the variance of features away from their respective class centers. φ is defined as:

$$\varphi = \frac{1}{2}\sum_{i=1}^{k}\left\| x_i - c_{y_i} \right\|_2^2 . \tag{8}$$

We used SGD technique to optimize the parameters. Parameter P was exploited to limit the objective number to the nearest class centers. The MD loss function in (6) is then improved as follows:

$$L_{\text{MD}} = -\sum_{i=1}^{N}\log\frac{\exp\left(\frac{1}{\varphi}\left\| x_i - c_{y_i} \right\|_2^2 - \lambda\right)}{\sum_{j=1,j\neq i}^{P}\exp\left(-\frac{1}{\varphi}\left\| x_i - c_{y_i} \right\|_2^2\right)} \tag{9}$$

where P is carefully selected to efficiently skip some class centers which are relative far away from the current center.

Given a set of images from N different datasets, where the label spaces are different, we define $L_{\text{ML}k}$ the multi-dataset loss of the kth dataset, the overall loss function is given:

$$L = \frac{1}{N}\sum_{i=1}^{N}\lambda_i L_{\text{MD}i} \tag{10}$$

where λ is the weight of each loss.

The update is performed in a mini-batch, which can avoid a large amount of calculation and increase systematic stability. Considering the classical BP algorithm, the entire parameter updating process of MD loss is summarized in Algorithm 1.

Algorithm 1 The parameter updating algorithm of MD loss.

Input: Training dataset $N\{x_i\}$, number of iterations T, learning rate η^t and α, and hyper-parameters P, λ_i.

Output: Network parameter W, model parameters θ, and the loss parameters c_j.

```
for $t = 1$ to $T$ do
    calculating function $L$ by:
    $L = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{1}{\varphi}\left\| x_i - c_{y_i} \right\|_2^2 - \lambda_i\right)}{\sum_{j=1,j\neq i}^{P}\exp\left(-\frac{1}{\varphi}\left\| x_i - c_{y_i} \right\|_2^2\right)}$
    updating the gradients by:
    $\theta^{t+1} = \theta^t - \mu\frac{\partial L^t}{\partial \theta^t}$
    updating the parameters by:
    $c_j^{t+1} = c_j^t - \alpha\Delta c_j^t$
    computing the BP error for each i by:
    $\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L_S^t}{\partial x_i^t} + \lambda\frac{\partial\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i L_{\text{MD}i}\right)^t}{\partial x_i^t}$
    updating the network parameters by:
    $W^p = \frac{\partial L^t}{\partial x_i^t}\frac{\partial x_i^t}{\partial W^t}$
    $W^{t+1} = W^t - \eta^t \cdot W^p$
end for
```

4. Experiment Results and Analysis

For deep feature learning, we employ the TensorFlow implementation, which is commonly used in several recent works. The first network was pre-trained by using the first dataset (i.e. the experimental facial emotion images) in MTE database as disputed in Sec. 3.1. All samples were divided into training data (70%) and test data (30%). The database learning rate is set to 0.01, which will be divided by 10 after every 10,000 iterations. In each iteration, 256 samples are used for stochastic gradient optimization. After 200 epoch's training, the first sub-network obtained 70.14% accuracy on the first dataset in MTE. The second network was pre-trained on the second data in MTE. The training strategy was same as the first network training. The second sub-network gets 64.11% accuracy on the MTE database, after 311 epoch's training. In the fine-tune stage, we exchanged the tuning datasets, that is, the second dataset for the first network and the first dataset (i.e. the experimental facial emotion dataset) for the second network. The base learning rate is changed to 0.001. The validation accuracy is converged after 200 epoch's fine-tuning.

Note that, those two Softmax functions were only used in the training step. All the experiments in this work were developed on NVIDIA GeForce GTX Titan GPU. We used dropout technique to each fully-connect layer with a probability of 0.65. Accuracy measures the number of correctly classified examples; it is defined as follows:

$$accuracy_i = \frac{TP_i + TN_i}{N} \tag{11}$$

where i species the class, i.e., the i-th emotion category, TP_i (true positives) are correctly identified test instances of class i, TN_i (true negatives) are test images correctly labeled as not belonging to class i, and N is the total number of test images.

Face detection: We use the Selective Search algorithm in a similar manner as RCNN to generate region proposals for faces in an image. A region having an overlap of more than 0:5 with the ground truth bounding box is considered a positive sample (l = 1). The candidate regions with overlap less than 0:35 are treated as negative instance (l = 0). All the other regions are ignored.

Face alignment: Face alignment is conducted to reduce variation in face scale and in-plane rotation across different facial images. In this work, we use the face alignment strategy proposed in [24] to robustly align human face 'in-the-wild'.

The next experiment is for emotion recognition which was performed on the MTE database. The confusion matrix of the proposed multitask CNN model is reported in Tab. 1.

We also test our approach on the well-known datasets which are available on public websites. The confusion matrix on CK+ [28] and MMI [29] are listed in Tab. 2 and Tab. 3, respectively. The CK+ database consists of 593 sequences from 123 subjects. The MMI database includes 30 subjects of both sexes and ages from 19 to 62. In the datasets, 213 sequences have been labeled with six basic expressions, in which 205 sequences are captured frontal view.

Emotion recognition experiment demonstrates that the proposed scheme is able to be used in real environment. Note that our approach performed unsatisfactorily in MMI dataset when recognizing the fear emotion. Anyway, few methods can perform perfectly in fear recognition since this emotion is too imperceptible to be learned in MMI.

	happy	sadness	surprise	anger	disgust	fear	neutral
happy	**83.4**	3	3	1.5	5.3	2	3
sadness	0	**82.6**	3.2	3	5	5	2.2
surprise	5.6	0	**86.3**	2	3.2	19.8	0
anger	1.3	2.5	3	**73.5**	8.6	5.5	2.2
disgust	5.5	8.9	0	6.6	**71.5**	3	5.8
fear	6.2	7.6	25.8	5.3	2.5	**79.2**	3
neutral	3	5.5	2.6	0	8.8	5.8	**81.5**

Tab. 1. Confusion matrix (in percentage) of the proposed method evaluated on the FER2013 database. The ground truth and the predicted labels are given by the first column and the first row, respectively.

	happy	sadness	surprise	anger	disgust	fear	contempt
happy	**100**	0	0	0	0	0	0
sadness	0	**93.5**	0	3.2	0	1.8	2.7
surprise	0	0	**98.5**	0.2	0	1.8	0
anger	0	7.5	1.2	**93.1**	0	0	0
disgust	0	0	0	0	**98.8**	0	0
fear	1.5	0	9.3	0	0	**96.8**	3.6
neutral	0	5.8	0	5.8	0	0	**88.5**

Tab. 2. Confusion matrix (in percentage) of the proposed method evaluated on the CK+ dataset. The ground truth and the predicted labels are given by the first column and the first row, respectively.

	happy	sadness	surprise	anger	disgust	fear	neutral
happy	**93.8**	2.8	3	0	2.4	0	2.4
sadness	0	**78.4**	3.2	9.5	6.1	0	7.5
surprise	5.6	0	**79.5**	8.1	0	12.1	6.7
anger	1.8	2.5	0	**82.2**	8.6	2.4	5.2
disgust	5.6	7.9	2.5	7.8	**71.2**	3.2	0.8
fear	4.1	5.9	5.5	4.8	1.2	**45.8**	5.4
neutral	0	2.5	0	4.7	0	0	**83.5**

Tab. 3. Confusion matrix (in percentage) of the proposed method evaluated on the MMI dataset. The ground truth and the predicted labels are given by the first column and the first row, respectively.

To evaluate the proposed approach comprehensively, we make a comparison between the proposed approach and some state-of-the-art FER methods [30–32]. The database used in this experiment was consisted with CK+ [28], MMI [33] and MTE. We used 70% of the total samples for training and the rest for testing. As shown in Tab. 4, experimental result demonstrates that our scheme obtains a better reorganization performance than others which is benefited from contribution of multi-dataset architecture and the MD loss function. The overall time taken to perform all the two tasks was 2 s per image. The results indicate that our approach has considerably higher performance. We also replaced the loss function in [30–32] with the proposed MD loss function to test the validity. Results are listed in the third column of Tab. 4 (the value behind the slash). The proposed MD loss function would increase the accuracy of FER by enhancing the discriminative power of the deeply learned features.

To further test the usability of our proposed multi-dataset network, we evaluated the performance of facial landmark localization. The algorithm proposed in [24] was used to extract the facial landmarks localization data in MTE database. The training and fine-tuning steps were the same as the FER task described at the beginning of this section. Some of the methods for comparison include Li et al. [34], Ramanan et al. [21] and Chen et al. [35].

Method	Accuracy %
Zhang et al. [30]	75.10/75.21
Kim et al. [31]	72.72/74.55
Meng et al. [32]	76.82/78.27
Proposed	79.7

Tab. 4. Performance comparison of different methods.

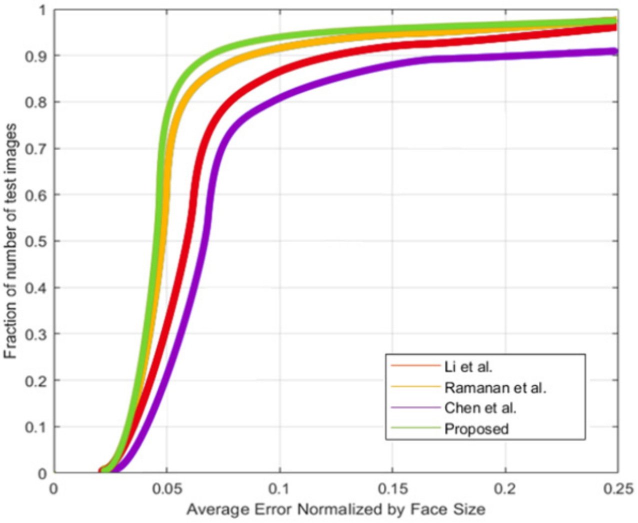

Fig. 6. Cumulative error distribution curves for landmark localization.

The comparison result shown in Fig. 6 demonstrates that our multi-dataset strategy can be perfectly applied to facial landmarks localization.

5. Conclusion

In this paper, we presented a novel facial expression recognition method based on machine learning technique. A neural network with multiple datasets was proposed to learn the expression features in different datasets, which can increase the robustness of FER task. To efficiently fuse the discriminative power of the deeply learned features, we proposed a novel multi-dataset loss function based on the center loss algorithm. Extensive experiments demonstrated that the proposed method is able to recognize both experimental facial emotion and facial expressions in wild scenarios. In future, we will evaluate the performance of our method on other applications such as simultaneously predict facial landmarks, human pose and human gender.

Acknowledgments

This work is supported in part by the Humanities and Social Sciences Projects of the Ministry of Education (NO. 18YJC760112); the Social Science Fund of Jiangsu Province (NO. 18YSD002); the Fundamental Research Funds for the Central University (NO. 2019JDZD02).

References

[1] PIANA, S., STAGLIANÒ, A., ODONE, F., et al. Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems,* 2016, vol. 6, no. 1, p. 1–31. DOI: 10.1145/2818740

[2] TIAN, Y., KANADE, T., COHN, J. F. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2001, vol. 23, no. 2, p. 97–115. DOI: 10.1109/34.908962

[3] LI, S., DENG, W. *Deep Facial Expression Recognition: A Survey.* 2018, p. 1–25. arXiv:1804.08348v2

[4] MOLLAHOSSEINI, A., CHAN, D., MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Placid (NY, USA), 2016, p. 1–10. DOI: 10.1109/WACV.2016.7477450

[5] JUNG, H., LEE, S., YIM, J., et al., Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision*. Santiago (Chile), 2015, p. 2983–2991. DOI: 10.1109/ICCV.2015.341

[6] FASEL, B. Head-pose invariant facial expression recognition using convolutional neural networks. In *Proceedings of the IEEE International Conference on Multimodal Interfaces*. Pittsburgh (PA, USA), 2002, p. 1–6. DOI: 10.1109/ICMI.2002.1167051

[7] MATSUGU, M., MORI, K., MITARI, Y., et al. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks,* 2003, vol. 16, no. 5, p. 555–559. DOI: 10.1016/S0893-6080(03)00115-1

[8] SIKKA, K., DYKSTRA, K., SATHYANARAYANA, S., et al., Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. Sydney (Australia), 2013, p. 517–524. DOI: 10.1145/2522848.2531741

[9] CHEN, J., CHEN, Z., CHI, Z., et al. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *International Conference on Multimodal Interaction*. Istanbul (Turkey), 2014, p. 508–513. DOI: 10.1145/2663204.2666277

[10] GIRSHICK, R., DONAHUE, J., DARRELL, T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. Columbus (OH, USA), 2013, p. 580–587. DOI: 10.1109/CVPR.2014.81

[11] SUN, B., LI, L., ZHOU, G., et al. Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging,* 2016, vol. 25, no. 6, p. 1–8. DOI: 10.1117/1.JEI.25.6.061407

[12] PONS, G., MASIP, D. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. 2018, p. 1–9. arXiv:1802.06664

[13] YANG, Y., HOSPEDALES, T. M. Unifying multi-domain multitask learning: Tensor and neural network perspectives. Chapter in CSURKA, G. (ed.) *Domain Adaptation in Computer Vision Applications,* 2017, p. 291–309. DOI: 10.1007/978-3-319-58347-1_16

[14] THRUN, S., PRATT, L. *Learning to Learn*. Boston, MA: Springer US, 1998. DOI: 10.1007/978-1-4615-5529-2

[15] FOURURE, D., EMONET, R., FROMONT, E., et al. Multi-task, multi-domain learning: Application to semantic segmentation and pose regression. *Neurocomputing,* 2017, vol. 251, no. C, p. 68–80. DOI: 10.1016/j.neucom.2017.04.014

[16] OBERWEGER, M., LEPETIT, V. DeepPrior++: Improving fast and accurate 3D hand pose estimation. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*. Venice (Italy), 2018. p. 585–594. DOI: 10.1109/ICCVW.2017.75

[17] ZHANG, K., ZHANG, Z., LI, Z., et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters,* 2016, vol. 23, no. 10, p. 1499–1503. DOI: 10.1109/LSP.2016.2603342

[18] XIAN, Y., HU, H. Enhanced multi-dataset transfer learning method for unsupervised person re-identification using co-training strategy. *IET Computer Vision,* 2018, vol. 12, no. 8, p. 1219–1227. DOI: 10.1049/iet-cvi.2018.5103

[19] RANJAN, R., PATEL, V. M., CHELLAPPA, R. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2018, vol. 41, no. 1, p. 121–135. DOI: 10.1109/TPAMI.2017.2781233

[20] GOODFELLOW, I., ERHAN, D., LUC CARRIER, P., et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks,* 2015, vol. 64, p. 59–63. DOI: 10.1016/j.neunet.2014.09.005

[21] DHALL, A., GOECKE, R., LUCEY, S., et al. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *IEEE International Conference on Computer Vision Workshops*. Barcelona (Spain), 2011, p. 2106–2112. DOI: 10.1109/ICCVW.2011.6130508

[22] ASTHANA, A., ZAFEIRIOU, S., CHENG, S., et al. Incremental face alignment in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. Columbus (OH, USA), 2014, p. 1859–1866. DOI: 10.1109/CVPR.2014.240

[23] NAIR, V., HINTON, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* Haifa (Israel), 2010, p. 807–814

[24] SHANG, W., SOHN, K., ALMEIDA, D., et al. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *Proceedings of the 33rd International Conference on Machine Learning*. New York (USA), 2016, p. 2217–2225.

[25] SZEGEDY, C., LIU, W., JIA, Y., et al. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston (MA, USA), 2015, p. 1–9. DOI: 10.1109/CVPR.2015.7298594

[26] WEN, Y., ZHANG, K., LI, Z., et al. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*. Munich (Germany), 2016, p. 499–515. DOI: 10.1007/978-3-319-46478-7_31

[27] SAKR, C., PATIL, A., ZHANG, S., et al. Minimum precision requirements for the SVM-SGD learning algorithm. In *IEEE International Conference on Acoustics*. New Orleans (LA, USA), 2017, p. 1138–1142. DOI: 10.1109/ICASSP.2017.7952334

[28] LUCEY, P., COHN, J. F., KANADE, T., et al. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. San Frencisco (CA, USA), 2010, p. 94–101. DOI: 10.1109/CVPRW.2010.5543262

[29] PANTIC, M., VALSTAR, M., RADEMAKER, R., et al. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*. Amsterdam (Netherlands), 2005, p. 5–13. DOI: 10.1109/ICME.2005.1521424

[30] ZHANG, Z., LUO, P., LOY, C. C., et al. Learning social relation traits from face images. In *IEEE International Conference on Computer Vision*. Santiago (Chile), 2015, p. 3631–3639. DOI: 10.1109/ICCV.2015.414

[31] KIM, B. K., ROH, J., DONG, S. Y., et al. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces,* 2016, vol. 10, no. 2, p. 173–189. DOI: 10.1007/s12193-015-0209-0

[32] MENG, Z., LIU, P., CAI, J., et al. Identity-aware convolutional neural network for facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*. Washington (DC, USA), 2017, p. 558–565. DOI: 10.1109/FG.2017.140

[33] LI, H., HUA, G., LIN, Z., et al. Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE International Conference on Computer Vision*. Sydney (Australia), 2014, p. 793–800. DOI: 10.1109/ICCV.2013.103

[34] CHEN, D., REN, S., WEI, Y., et al. Joint cascade face detection and alignment. In *European Conference on Computer Vision*. Zurich (Switzerland), 2014, p. 109–122. DOI: 10.1007/978-3-319-10599-4_8

About the Authors ...

Bin YANG was born in Shaoguan City, in 1979. He received his M.S. degree in Computer Science from the South China University of Technology, China, in 2007, and Ph.D. in Computing Science and Technology from the Hunan University, China, in 2014. From 2002 to 2010, he was a lecturer in the South China Normal University. Since 2014, he has been an Associate Professor at the Jiangnan University, China. His research interests include information security, digital image forensic, machine learning and image processing. He has authored over 30 papers in related areas.

Zhenyu LI received his M.S. and Ph.D. degrees from Sangmyung University, Seoul, Korea, in 2011 and 2015, respectively. Since 2015, he has been a lecturer at the Jiangnan University, Wuxi, China. His research interests include machine learning, multimedia processing, and information security.

Enguo CAO received his M. Eng. degree in Mechanical Engineering from Yanshan University in China in 2009, Ph.D. degree in Intelligent Mechanical Systems Engineering from Kochi University of Technology in Japan in 2012. He was a Mechanical engineer in the National secondary heavy machinery group of China in 2009, and he was a Research Assistant in the Dep. of Intelligent Mechanical Systems Engineering, Kochi University of Technology, Kochi, Japan. He is currently an Associate Professor, head of intelligent interaction design lab of the Jiangnan University in China. His research interests include Intelligent Interaction System and Rehabilitation Product Design.